RegExp in Java, TOS and IntelliJ

Arkady Galyash

TosChart

May 25, 2012

Agenda

1. RegExp in Java
2. RegExp in TOS
3. RegExp in IntelliJ
4. Links

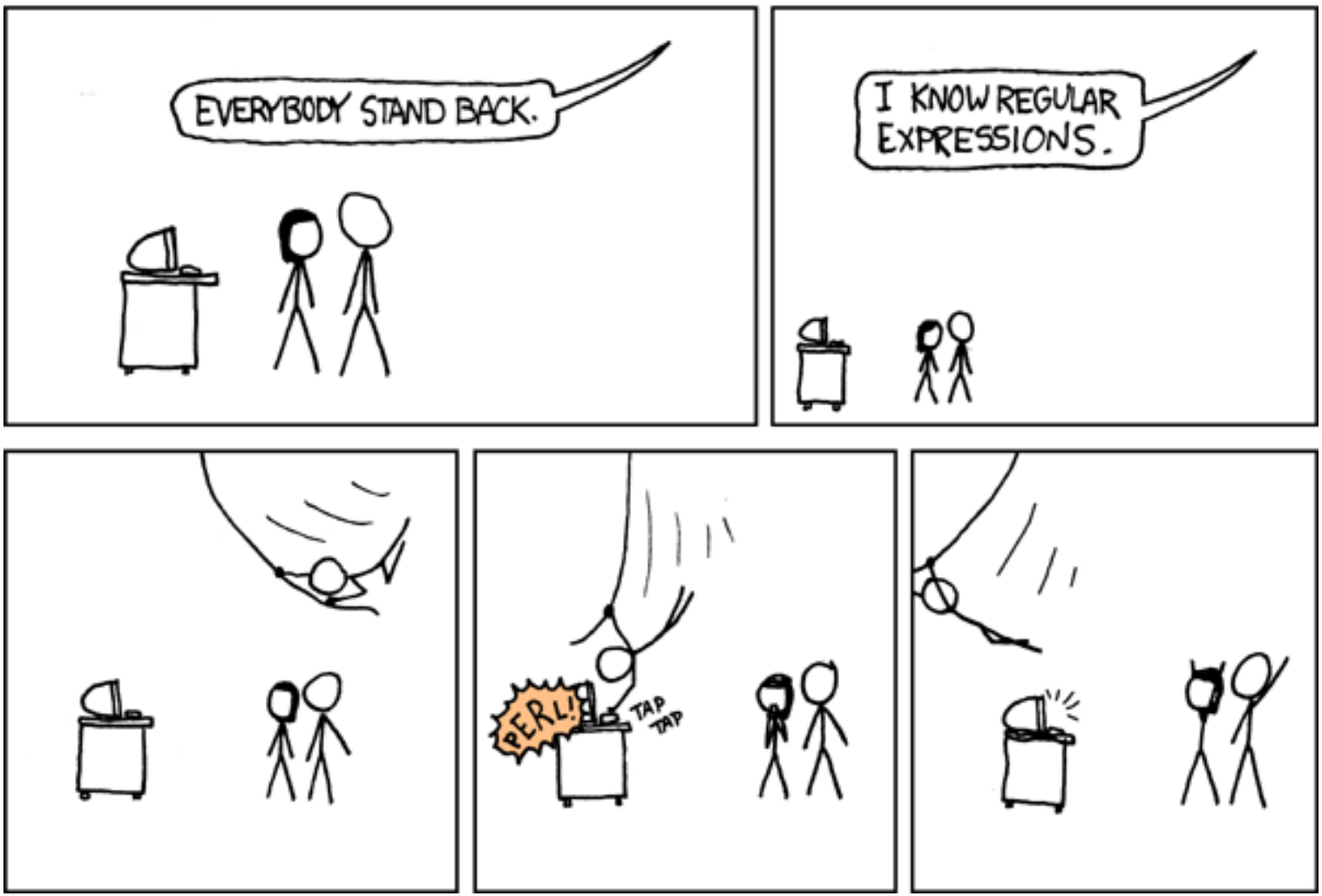
History

- Before UNIX(Stephen Cole Kleene, 1950s): automata theory
- UNIX(Ken Thompson, 1970s): QED, ed, g/re/p, ...
- Perl(Larry Wall, 1988): 2nd edition

Why developer should know RegExp?

- Learn once, use anywhere
- Extract needed info from log files(grep)
- Advanced text editing
- Using lexer generators(lex) and syntax highlighting
- ...

Do you want to be a Superhero?



With great power there must also come great responsibility!

— Stan Lee, “Spider-Man”

Some people, when confronted with a problem, think ”I know, I’ll use regular expressions.” Now they have two problems.

— Jamie Zawinski, alt.religion.emacs

RegExp in Java

- simple strings, compiled by java.util.regex.Pattern
- no qr//, concatination of strings is really bad idea
- “backslashes hell”

RegExp CodeStyle in TOS

- Real task.
- Used solution(case).
- Guess what's wrong with this case codestyle?
- Thoughts about this case and how to avoid wrong patterns.
- Discussion.

RegExp #1

Task:	text inside {}
Case:	`(\\{([^\\}]*)\\})`
Fix:	`(\\{([^}]*)\\})`
Rule:	Do not escape characters that do not need escaping inside character classes

RegExp #2

Task:	match whitespace characters(blank, CR, NL, TAB, etc.)
Case:	`[\\s\r\n]*`
Fix:	`[\\s]*`
Rule:	Remove from character classes single chars that are already there

RegExp #3

Task:	semicolon with leading whitespaces
Case:	`[\\s]*[;]`
Fix:	`\\s*;`
Rule:	Remove character classes when they are unneeded

RegExp #4

Task:	part of complex regexp
Case:	`(?i)(([\\n\\r]+)|(^))`
Fix:	`(([\\n\\r]+)|(^))`
Rule:	Remove useless inline modifiers

java.util.regex.Pattern#compile(...)

```
String A = "(?i)\s*";
String B = "test";
...
Pattern p = Pattern.compile(A + B);
```

VS

```
String A = "\s*";
String B = "test";
...
Pattern p = Pattern.compile(A + B,
                Pattern.CASE_INSENSITIVE);
```

RegExp #5

Task:	part of complex regexp #2
Case:	`(([\n\r]+)|(^))`
Fix:	`(([\n\r]+)|^)`
Rule:	Remove useless capturing(for zero-width assertions)

RegExp #6

Task:	copyright
Case:	`\\(company\\.  \\(c\\)\\)`
Fix:	`\\Q(company.  (c))\\E`
Rule:	Use \Q...\E when it is appropriate

RegExp HowTo #1

Task: digit

Cases: `[0-9]` or `\\d`

Prefer: `\\d`

RegExp HowTo #2

Task: any symbol(including NL, CR, etc.)

Cases: `[\\w\\W]` or `(?s:.)`

Prefer: `(?s:.)`

RegExp HowTo #3

Task: How to escape special chars?

Cases: `[.]` or `\\.`

Prefer: `[.]`

RegExp HowTo #4

Task: CR, Tab, NL

Cases: `\\n` or `\n`

Prefer: `\n`

RegExp inspections

- 10 inspections(old-fashion, work in CE too)
- 2 weekends
- 2000 LOC in 25 *.java files(including unit tests)
- A lot of kudos to Sascha Weinreuter(language injection plugin)

RegExpNazi

https://github.com/gark87/RegExpNazi

Start

- package com.intellij.codeInspection
- plugin.xml:
  <inspectionToolProvider implementation="..."/>
- interface InspectionToolProvider
  implement Class[] #getInspectionClasses()
- LocalInspectionTool

LocalInspectionTool

- ProblemDescriptor[] #checkFile(...)
  check only our filetypes
- String #getGroupDisplayName()
  String #getDisplayName()
- String #getShortName()
  HTML file with help
- JComponent #createOptionsPanel()
  use public fields to store options

Miscellaneous

- LocalQuickFix
  - #applyFix()
  - fix AST and PSI trees properly
  - if replacement is really hard to construct then you can use fake file
- Unit tests
  - file-based
  - -Didea.platform.prefix=”Idea”

Links

- coursera.org course by Jeffrey Ullman
- Regular Expression Matching Can Be Simple And Fast(but is slow in Java, Perl, PHP, Python, Ruby, ...) by Russ Cox
- “Mastering Regular Expressions” by Jeffrey Friedl
- @RegexTip
- man perlre*

Thank You!